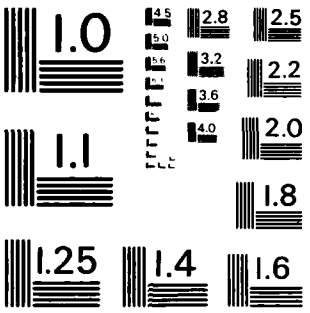
AD-A135 850 RESEARCH ON NONPARAMETRIC METHODS FOR LINEAR MODELS(U) 1/1
WESTERN MICHIGAN UNIV KALAMAZOO DEPT OF MATHEMATICS
G L SIEVERS NOV 83 TR-73 N00014-78-C-0637

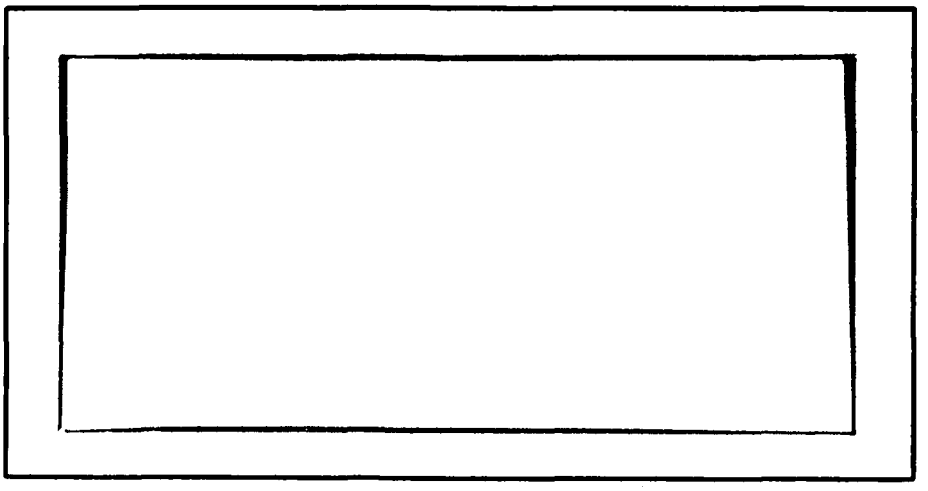
UNCLASSIFIED F/G 12/1 NL

END
DATE FILMED
01-84
DTIC



(17)

AD-A135850



DEPARTMENT OF MATHEMATICS

WESTERN MICHIGAN UNIVERSITY

Kalamazoo, Michigan

49008

DTIC FILE COPY

83 12 14 017

17

FINAL REPORT

ON THE PROJECT

"RESEARCH ON NONPARAMETRIC METHODS FOR LINEAR MODELS"

by

Gerald L. Sievers

TECHNICAL REPORT NO. 73
November 1983

This research was supported by the Office of Naval Research under contract N00014-78-C-0637 (NR 042-407)

DEPARTMENT OF MATHEMATICS
WESTERN MICHIGAN UNIVERSITY
KALAMAZOO, MICHIGAN 49008

A-1

THE GENERAL RESEARCH PROJECT

The goal of this project was to conduct research in the development of nonparametric statistical methods based on rank statistics for the general linear model. Consideration was to be given to the use of a comprehensive framework to handle estimation, confidence regions and tests of hypotheses for a wide range of multiple regression and analysis of variance problems. Methods were sought which would be efficient and convenient to use in practice. The practical aspects of implementation for applied problems requires special attention in order that the methodology be direct in interpretation and computationally feasible.

The general linear model can be represented by the equation $y = Xb + e$, where y is an nxl vector of observations on a dependent variable, X is an nxp design matrix, b is a pxl vector of unknown parameters and e is an nxl vector of independent random errors with common density $f(y)$. A general procedure for obtaining an estimate of b is to minimize some measure of dispersion of the residuals $Z = Y - Xb$. Least squares estimation, M - estimation and rank estimation follow this plan. In the present work the estimate $\hat{b}$ is the point minimizing a dispersion function $D(b) = \Sigma w_{ij}|Z_i - Z_j|$, where the w_{ij} are a set of weights. The derivative $T(b)$ of the dispersion function is a vector of weighted rank statistics. A further generalization of interest is the dispersion function $\Sigma w_{ij}\, p(Z_i, Z_j)$, where p is a suitably chosen function.

The special case where all weights equal one in the dispersion function is a benchmark case and the estimate $\hat{b}$ becomes the reg-

ular rank estimate based on Wilcoxon scores. In general, the properties of $\hat{b}$ and other statistics based on D(b) or T(b) depend on the weights used. The use of zero weights can drop some pairs of residuals from consideration and restricted rank statistics result. Low weights can be used to diminish the effects of outliers and high leverage points in the design. The weights can be chosen to match the particular design matrix on hand. An important problem is to identify weights that would result in no loss of efficiency.

In order to determine the large sample behavior of $\hat{b}$ and the effect of the weights on this estimate the asymptotic distributions of D(b) and T(b) are needed. These asymptotic results are also needed in studying the properties of test statistics and their asymptotic efficiencies.

RESEARCH ACCOMPLISHMENTS

1. Sievers, G.L. (1979), "A Weighted Dispersion Function For Estimation in Linear Models," Technical Report No. 62, Department of Mathematics, Western Michigan University. Also an expanded version under the same title appeared in 1983, Communications in Statistics, Theory and Methods - A, Vol. 12, p. 1161 - 1179.

This report develops the basic asymptotic distribution theory for the estimate $\hat{b}$ minimizing the dispersion function D(b). Equivalently, $\hat{b}$ is a zero of the weighted rank vector T(b). The vector T(b) is shown to be asymptotically multivariate normal. An asymptotic linearity result is proven for T(b) and this shows the dispersion function is locally quadratic. These results lead to the asymptotic normality of $\hat{b}$.

An influence function is developed and it shows that the weights can be used to lessen the effect of high leverage points in the design variables. It is shown that the use of weights cannot improve the efficiency of the estimate over the unweighted case but a condition is given under which there would be no loss of efficiency with the use of weights. This is illustrated in several examples.

2. Sievers, G.L. and Kapenga, J. (1981), "Approximate Empirical Distributions For the Computation of Nonparametric Statistics," Technical Report No. 64, Department of Mathematics, Western Michigan University.

This report was motivated by computational concerns at this stage of the project. It presents a method of approximating the value of a statistic through the use of a grouped frequency distribution with bounds on the error of approximation. This is of particular interest for computing rank statistics and other nonparametric statistics since it avoids the time consuming process of ordering observations. An approximation to a rank statistic based on n observations can be computed in linear time, proportional to n, whereas the time for the exact calculation is of order n^2 or $n \log n$, depending on the sorting algorithm used. This savings in time can be especially useful with the iterative computations needed in minimizing the dispersion function since it avoids the sorting of residuals normally required at each step.

The approximation method can be adapted to attain any preassigned degree of accuracy. The approximations of the Wilcoxon signed rank statistic and the median of the Walsh averages are used

for illustration in the report. A simulation study is presented to compare the execution times of the approximation method against that of the bubble sort and the quick sort methods.

3. Sievers, G.L. (1982), "A Consistent Estimate of a Nonparametric Scale Parameter, " Institute of Statistics Mimeo Series #1501, Department of Statistics, The University of North Carolina.

A consistent estimate is presented for the scale parameter $g = \int f^2$. This parameter arises in the asymptotic covariance matrix of $\hat{b}$ and it must be estimated in order to provide standard errors for the estimate. It also arises in the hypothesis testing problem as a scale factor for the test statistic based on the difference in the dispersion function between the reduced and the full model. The proposed estimate does not require the assumption of symmetry of the underlying density of the error variables. The previous Lehmann type estimate had required this symmetry.

Two versions of the scale estimate are discussed; one is related to a window estimate of a density function and the other to a nearest neighbor type estimate. The theory here is considerably complicated by the fact that the scale estimate must be based on residuals rather than on independent, identically distributed observations.

The scale estimate as developed here involves weights assigned to each pair of residuals. In analysis of variance problems the use of zero weights for pairs of residuals arising from different cells makes the scale estimate depend only on the within-cell variation and as a result it is independent of the parameter estimate $\hat{b}$.

4. Sievers, G.L. (1983), "A Robust Multiple Correlation Coefficient For the Rank Analysis of Linear Models," Technical Report No. 69, Department of Mathematics, Western Michigan University.

The multiple correlation coefficient R^2 is widely used in the analysis of linear models as a measure of the degree of association between a random variable Y and a set of random variables $X_1, \ldots, X_p$. However, it lacks robustness and can be sensitive to outliers. This report develops a robust multiple correlation measure defined in terms of a weighted Kendall's tau, suitably normalized. This new statistic is directly compatible with the rank statistic approach of analyzing linear models in a regression, prediction context since it measures the association between the Y observations and the fitted values from the linear model.

There is an underlying population parameter for the robust multiple correlation coefficient that equals the classical parameter when the multivariate normal model holds. For general distributions this parameter has several desireable properties.

The report proves that the sample statistic is a consistent estimate of the population parameter. The theory uses the basic results of report #1 and develops some new asymptotic theory for the fitted values of the rank analysis to obtain this result. A test of independence is discussed in this context.

5. Sievers, G.L. and McKean, J.W. (1983), "On the Robust Analysis of Linear Models With Nonsymmetric Error Distributions," Technical Report No. 70, Department of Mathematics, Western Michigan University.

This report is a general summary of the rank methodology for the rank analysis of linear models in the case of nonsymmetric error distributions. The emphasis is on the hypothesis testing problem. The results are also valid for symmetric distributions. The new methodology is based on the estimate of the scale factor developed in report #3 which did not require the symmetry assumption. A Monte Carlo study is presented which compares the performance of tests of hypotheses based on least squares, on the old rank analysis and on this new rank analysis for three small sample designs. The new rank analysis proved to be far superior to least squares in maintaining significance level and in power over a range of error distributions. It was slightly superior to the old rank analysis.

The Monte Carlo results in this report were a small part of a larger study on the performance of the rank methods in small sample linear model problems. This project will be continued and results prepared later. The work here has suggested that modifications to allow for skewed scores will improve the efficiency of the tests and this point will be studied further.

6. Sievers, G.L. (1983), "Testing Hypotheses in Linear Models With Weighted Rank Statistics," Technical Report No. 71, Department of Mathematics, Western Michigan University.

This report develops the general hypothesis testing theory and methodology for the weighted rank approach to the analysis of linear models. It continues the work of report #1 which had focused on the estimation problem. The main problem considered is to test a subhypothesis of the linear model which contains nuissance parameters.

Three types of tests are discussed: tests based on the estimate, tests based on the difference of the dispersion function D(b) between a reduced model and the full model and quadratic form tests based on the aligned rank procedure.

The weights used in the rank tests here are not designed to increase efficiency. Rather weights can be chosen to retain the highest efficiency and yet gain in other respects. The report contains details and examples on the practical application of this in several types of analysis of variance problems. For instance, nuissance parameters can be eliminated by restricting the ranking to within-block comparisons in block designs. In testing ordered alternatives, the weights can be used to restrict the comparisons to only adjacent groups. In factorial experiments the effects of model inadequacies can be diminished by restricitng the ranking to neighboring cells with the use of zero weights. In a sense the work here provides a common framework for many results that have been obtained separately in the literature and shows how modifications and extensions can be made in the same general framework.

7. Sievers, G.L. (1983), "Robust Estimation Based on Walsh Averages For the General Linear Model," Technical Report No. 72, Department of Mathematics, Western Michigan University.

This report develops the theory and asymptotic distribution of the estimate $\hat{b}$ obtained by minimizing a dispersion function of the form $\sum w_{ij}\, p(Z_i + Z_j)$ for a suitable convex function p. When the function p is the absolute value function this approach will generate weighted signed rank statistics in a manner similar

to the previous weighted rank statistics. Other p functions can be used to generate weighted or trimmed Walsh average statistics. The methods are very similar to the M-estimation approach but applied to Walsh averages. Details are discussed for several choices of p function.

The emphasis in this report is on the theory of the estimation problem. Tests of hypotheses could be developed following the work in report #6. The practical value and details of implementation in particular cases needs further study.

The work of this report, with slight modification, can yield the corresponding results for dispersion functions of the form $\sum w_{ij}\, p(Z_i - Z_j)$. This would be a generalization of the dispersion function $D(b)$ used in the earlier work.

COMPUTER PROGRAMMING

Considerable effort has been devoted to the problem of writing a computer program to carry out the computations for the methodology of this project. The heart of the problem is to find an accurate, efficient, iterative algorithm for the minimization of the diversion function $D(b)$. $D(b)$ is a well-behaved function, being convex and piecewise linear, but it is a sum of $\binom{n}{2}$ terms and iterative computations can become quite excessive even for moderate sample size n. This is a general problem for all analyses based on ranks. Methods such as steepest descent, Newton's method, Marquardt's compromise, variable metric and a modified dogleg were tried on problems up to size $n = 100$, $p = 20$. Although these methods worked, the time required was not acceptable.

An algorithm has been developed that takes into account the specific nature of the dispersion function. It performs much better than the general numerical algorithms and can handle larger sized problems in a reasonable time. It is based on the local asymptotic linearity of the gradient vector (see Report #1) and could be described as a modified Newton's Method. The algorithm is globally convergent and its rate of convergence is in between the linear rate of steepest descent and the quadratic rate of Newton's Method, depending on how closely the assumptions behind asymptotic linearity may hold. There are several versions of the algorithm differing by the way a scale factor is handled.

The developmental programs and algorithms devised for minimizing the dispersion function will be a key ingredient in the final computer program. Work is still continuing on this project and it is expected that a comprehensive program for the robust analysis of a linear model will be released in the public domain.

ACKNOWLEDGEMENT

The principal investigator is grateful for the fine contributions of John Kapenga on programming and general aspects of this project over several summers. Lee Witt and Daniel Cheung also assisted on this project in the spring of 1983.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Technical Report #73	2. GOVT ACCESSION NO. AD-A125 956	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Final Report on the Project "Research on Nonparametric Methods for Linear Models:		5. TYPE OF REPORT & PERIOD COVERED Technical Report 1978-1983
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Gerald L. Sievers		8. CONTRACT OR GRANT NUMBER(s) N 00014-78-C-0637
9. PERFORMING ORGANIZATION NAME AND ADDRESS Western Michigan University Kalamazoo, Michigan 49008		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR 042-407
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Statistics and Probability Program		12. REPORT DATE November 1983
		13. NUMBER OF PAGES 10
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for Public Release: Distribution Unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

None

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This is a final report on the work completed for the project "Research on Nonparametric Methods for Linear Models".

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)